Summary

A Bayesian Network to model the influence of some behavioural risk factors on cardiovascular diseases

Zarmina Ursino, zarmina.ursino@studio.unibo.it
Fundamentals of Artificial Intelligence and Knowledge Representation (Mod. 3) –
Alma Mater Studiorum Università di Bologna

Abstract

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. The most important behavioural risk factors of heart disease are physical inactivity, tobacco use and harmful use of alcohol. Their effects may show up in individuals as raised blood pressure, raised blood Glucose, and raised blood Cholesterol. These factors indicate an increased risk of heart attack, stroke, heart failure and other complications. In this project I'm going to clean the dataset, to analyse the Bayesian model by studying the independence assertions, the active trails and the Markov Blanket and I'm going to come to some conclusions thanks to the Variable Elimination Algorithm.

1. Introduction

Probabilistic Graphical Models (PGM) are a very solid way of representing joint probability distributions on a set of variables and it allows users to do inferences in a computationally efficient way. I am going to use the 'pgmpy' library to create and analyse the Bayesian Network.

2. Reading CSV file

A dataset of 70000 rows and 13 columns was considered [1]. Attributes' information is given below. There are 3 types of input features: 'Objective', which designates factual information, 'Examination' which indicates that values are results of medical examination and 'Subjective', which indicates information given by the patient. The features are:

- Age | Objective Feature | age | int (days)|
- Height | Objective Feature | height | int (cm)
- Weight | Objective Feature | weight | float (kg) |
- Gender | Objective Feature | Gender | categorical code |
- Systolic blood pressure | Examination Feature | Ap_hi | int |
- Diastolic blood pressure | Examination Feature | Ap_lo | int |
- Cholesterol | Examination Feature | Cholesterol | 1: normal, 2: above normal, 3: well above normal |
- Glucose | Examination Feature | Gluc | 1: normal, 2: above normal, 3: well above normal |
- Smoking | Subjective Feature | Smoke | binary |
- Alcohol intake | Subjective Feature | Alco | binary |
- Physical activity | Subjective Feature | Active | binary |
- Presence or absence of cardiovascular disease | Target Variable | Cardio | binary |

3. Data cleaning

I decide to remove the 'id' feature because it doesn't contribute to our target variable to achieve better a good Bayesian Network. I capitalize the first letter of the columns' names for the sake of notational conventions. Moreover, I apply some transformations, for example, I convert the patient's age into years rather than days and height into metres rather than centimetres. I group by period the ages, since 'pgmpy' doesn't support learning from continuous variables and so I categorize the columns in 'Infant', 'Toddler', 'Kid', 'Teen', 'Young Adult', 'Middle-aged Adult', 'Old Adult'. Then I get a better weight assignment by considering the BMI ('Body Mass Index'), proposed for the first time by the Belgian Adolphe Quetelet (1796-1874). By solving a formula that requires two known values, height and weight, the BMI calculation provides a coefficient to be entered into a special evaluation grid that allows the following to be established: 'normal' weight, 'underweight', 'overweight' and 'obesity' (the latter possibly classified into different levels of severity) [Figure 3.1]. Once this index is calculated, I remove the columns 'Weight' and 'Height' so that only BMI is considered.

$$\text{BMI}=\frac{weight\ in\ Kg}{(height\ in\ m)^2}$$

Blood pressure, it’s expressed as a measurement with two numbers, with one number on top (systolic) and one on the bottom (diastolic), like a fraction. The top number refers to the amount of pressure in the arteries during the contraction of the heart muscle. This is called systolic pressure ('Ap_hi'). The bottom number refers to your blood pressure when the heart muscle is between beats. This is called diastolic pressure ('Ap_lo'). The AHA [2] recommends specific ranges for healthy adults [Figure 3.2].

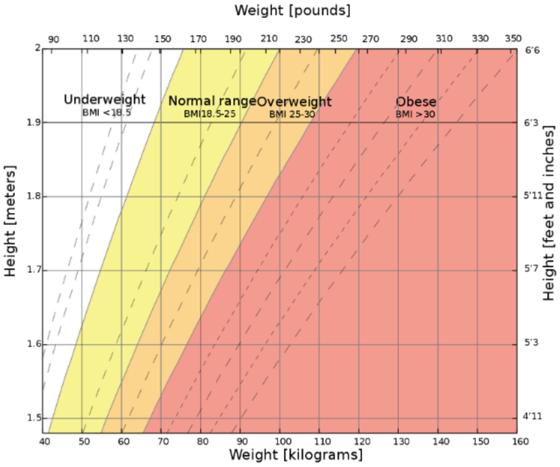
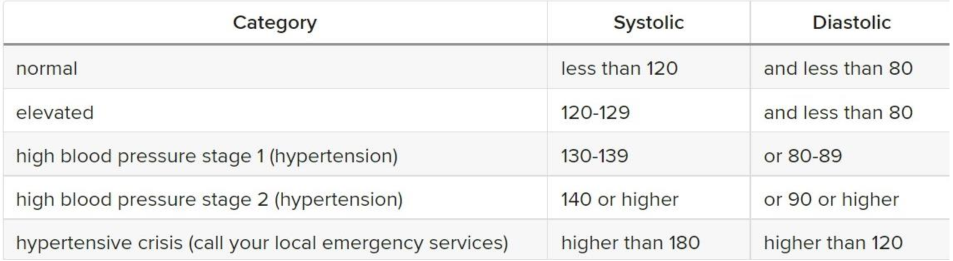


Figure 3.1 BMI chart



Category	Systolic	Diastolic
normal	less than 120	and less than 80
elevated	120-129	and less than 80
high blood pressure stage 1 (hypertension)	130-139	or 80-89
high blood pressure stage 2 (hypertension)	140 or higher	or 90 or higher
hypertensive crisis (call your local emergency services)	higher than 180	higher than 120

Figure 3.2 Systolic and Diastolic Pressure - Table

In fact, I categorise the ‘Ap_hi’ column in 'normal', 'elevated', 'hight_blood_pressure_stage1', 'hight_blood_pressure_stage2', 'hypertensive crisis'. On the other hand, I categorise the ‘Ap_lo’ in 'normal', 'hight_blood_pressure_stage1', 'hight_blood_pressure_stage2', 'hypertensive crisis'. Regarding the diastolic pressure, I do not consider the category 'elevated', as there is no variation between this and normal, as there is for systolic.

4. Bayesian Model

I draw the Bayesian Model [Figure 4.1] . The CPT is learned, by using a Bayesian Estimator, instead of a Maximum Likelihood Estimator, as it’s more robust. Then the CPT are displayed.

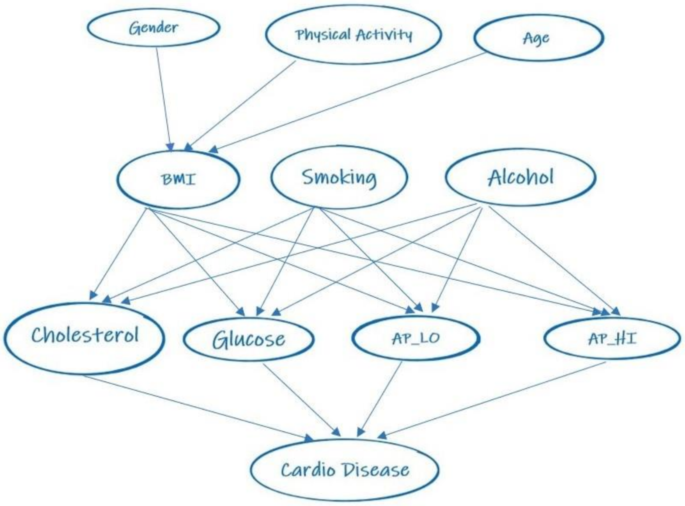


Figure 4.1 Bayesian Network

5. Independence

Thanks to the pgmpy method ‘get_independencies()’ [5] and ‘get_assertions()’ I print out the number of valid independence assertions with respect to all possible evidence .

6. Independence assertions

The independence assertions are implemented in pgmpy with the class 'IndependenceAssertion' [6], used to represent individual assertions of the form (X ⊥ Y) or (X ⊥ Y| Z). Thanks to the 'calculations()' function I evaluate which are the nodes that have the maximum and the minimum number of appearance in independence assertions as independent variable or as evidence.
IndependenceAssertion has 3 fields:
- 'event1', that is the variable which is independent.
- 'event2', variables from which event1 is independent.
- 'event3', variables given which event1 is independent of event2.

It can be notice that the closer a node is to the core of the network, the less are the independence assertions in which it is the independent variable (in fact BMI node appears 56 times in independence assertions as independent variable) and the more are the ones in which it is given as evidence (BMI node appears 4697 times in independence assertions as evidence). Moreover, the 'check_assertion()' function is implemented to propose some independence queries.

7. Active trails and Markov Blanket

In pgmpy it's possible to investigate different properties of the network. In fact, I implemented the functions 'active_trails_of()' [3], which returns all the nodes reachable from start via an active trail, and 'markov_blanket()' [4], which returns the set of node's parents, its children and its children's other parents.

8. Variable Elimination

The variable elimination algorithm acts on a set of factor, which involves a set of variables. Each node in a Bayesian network is equipped with a conditional probability function that expresses the likelihood that the node will take on different values given the values of its parents. Thus, given the probabilistic model created, I am interested in using this algorithm to answer questions that are useful in establishing an individual's risk of cardiovascular disease. Let's analyse some queries:

- Query 1: P(Cardio|alc=1, Gluc=3, chol=3, BMI=obese). As can be seen in the output, the probability of having a cardiovascular disease increases significantly if the patient has very high Glucose and Cholesterol values, drinks alcohol and has a BMI equal to obesity.
- Query 2: P(Cardio|alc=0, Gluc=2, chol=2, BMI=obese). In this case, the probability of contracting the disease decreases but is still very high as the BMI is still equal to obesity, even if Glucose and Cholesterol values have decreased slightly.
- Query 3: P(Cardio|alc=0, Gluc=1, chol=1, BMI=overweight). The probability of cardiovascular disease reduces if the individual does not drink alcohol, has BMI equal to overweight and has normal Glucose and Cholesterol levels.
- Query 4: P(Cardio|alc=0, Gluc=1, chol=1, BMI= normal weight). Here, the likelihood of having cardiovascular disease is greatly reduced, knowing that the individual does not drink alcohol, has normal Glucose and Cholesterol levels, and knowing that he or she has a normal BMI.
- Query 5: P(Cardio|alc=1, Gluc=3, chol=3, BMI= underweight). The result shows that overweight and obesity are not the only conditions that can cause cardiovascular disorders.
- Query 6: P(Cardio|alc=0, Gluc=1, chol=1, age=Young Adult). The probability of having CVDs is almost %50.
- Query 7: P(Cardio|alc=0, Gluc=1, chol=1, age=Old Adult). Lastly, the likelihood of having CVDs still remains almost %50. Thus, age does not significantly influence the probability of contracting the disease.

From the analysis of the Bayesian Network and especially from these queries we can conclude that cardiovascular diseases are generally caused by an unhealthy lifestyle characterized by smoking, drinking alcohol and a lack of physical activity (the latter influences the BMI) and not by the Age .

References

[1] Kaggle dataset, https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
[2] AHA, American Heart Association, https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings
[3] 'pgmpy' Documentation, https://pgmpy.org/models/naive.html
[4] 'pgmpy' Documentation, https://pgmpy.org/models/markovnetwork.html
[5] 'pgmpy' Documentation, https://pgmpy.org/base.html
[6] 'pgmpy' Documentation, https://pgmpy.readthedocs.io/en/latest/_modules/pgmpy/independencies/Independencies.html